(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **09319391 A**

(43) Date of publication of application: **12.12.97**

(51) Int. Cl **G10L 3/00**
G10L 5/04

(21) Application number: **08250150**

(22) Date of filing: **20.09.96**

(30) Priority: **12.03.96 JP 08 54714**
29.03.96 JP 08 77393

(71) Applicant: **TOSHIBA CORP**

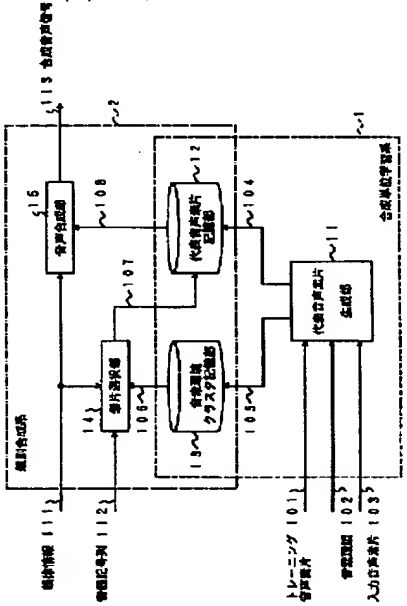
(72) Inventor: **KAGOSHIMA TAKEHIKO**
AKAMINE MASAMI

(54) SPEECH SYNTHESIZING METHOD

(57) Abstract:

PROBLEM TO BE SOLVED: To make it possible to effectively improve the tone quality of a synthesized speech by a text speech synthesis by forming such a representative phoneme that diminishes the distortion to natural speech at the level of the synthesized speech by taking the influence of change in pitch and the continuation time duration into consideration and synthesizing the speech by using the representative phoneme.

SOLUTION: The plural representative phonemes are formed by changing the pitch period and continuation time duration in a representative phoneme forming section 11. The distortion to the natural speeches is evaluated at the level of the synthesized speeches formed by changing at least either of the pitch and continuation time duration with respect to input phonemes 103. The phonemes selected from the input phonemes 103 based thereon are determined as the representative phonemes 104. The speech synthesis is executed in a speech synthesizing section 15 by connecting these representative phonemes 104, by which the high-quality synthesized speech signals 113 approximate to the natural speeches are formed.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－319391

(43)公開日　平成９年(1997)12月12日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
G10L 3/00			G10L 3/00	H	
5/04			5/04	F	

審査請求　未請求　請求項の数12　ＯＬ　（全25頁）

(21)出願番号　特願平8－250150

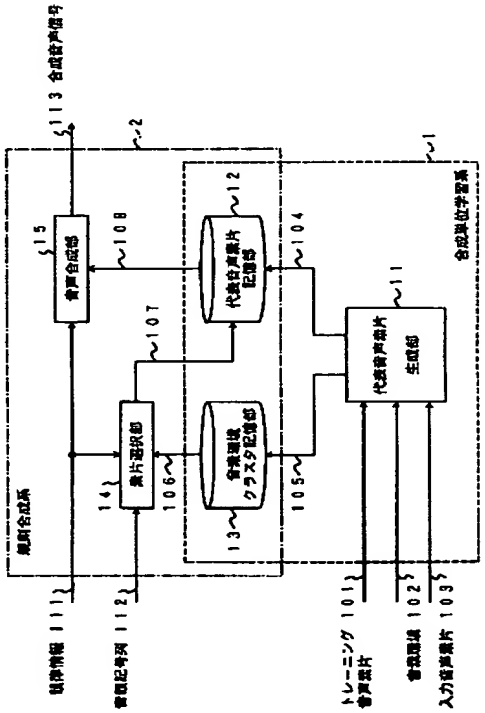
(22)出願日　平成８年(1996)９月20日

(31)優先権主張番号　特願平8－54714
(32)優先日　平８(1996)３月12日
(33)優先権主張国　日本（ＪＰ）
(31)優先権主張番号　特願平8－77393
(32)優先日　平８(1996)３月29日
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　籠嶋　岳彦
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　赤嶺　政巳
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】音声合成方法

(57)【要約】
【課題】テキスト音声合成による合成音声の音質を効果的に向上させることができる音声合成方法を提供する。
【解決手段】代表音声素片生成部１１において音素環境１０２がラベル付けされたトレーニング音声素片１０１のピッチ・継続時間長に従って入力音声素片１０３のピッチ・継続時間長を変更して複数の合成音声素片を生成し、合成音声素片とトレーニング音声素片１０１との間の距離尺度に基づいて入力音声素片１０３から代表音声素片１０４を選択して代表音声素片記憶部１２に記憶し、さらに距離尺度に基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタ１０５を生成して音素環境クラスタ記憶部１３に記憶し、代表音声素片記憶部１２から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を読み出して音声合成部１５で接続することにより、合成音声信号１１３を生成する。



1

【特許請求の範囲】

【請求項１】複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、
これらの合成音声素片と前記第１の音声素片との間の距離尺度に基づいて前記第２の音声素片から複数の代表音声素片を選択して記憶し、
これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【請求項２】音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、
これらの合成音声素片と前記第１の音声素片との間の距離尺度に基づいて前記第２の音声素片から複数の代表音声素片を選択して記憶し、
前記距離尺度に基づいて前記代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、
前記代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【請求項３】音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、
これらの合成音声素片と前記第１の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、
前記距離尺度に基づいて前記第２の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、
これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【請求項４】音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、
これらの合成音声素片と前記第１の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、
前記距離尺度に基づいて前記第２の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、
これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続

2

することによって音声を合成することを特徴とする音声合成方法。

【請求項５】複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、
これらの合成音声素片についてスペクトル整形を行い、
このスペクトル整形を行った後の各合成音声素片と前記第１の音声素片との間の距離尺度に基づいて前記第２の音声素片から複数の代表音声素片を選択して記憶し、
これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、
この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項６】音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、
これらの合成音声素片についてスペクトル整形を行い、
このスペクトル整形を行った後の各合成音声素片と前記第１の音声素片との間の距離尺度に基づいて前記第２の音声素片から複数の代表音声素片を選択して記憶し、
前記距離尺度に基づいて前記代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、
前記代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成し、
この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項７】音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、
これらの合成音声素片についてスペクトル整形を行い、
このスペクトル整形を行った後の各合成音声素片と前記第１の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、
前記距離尺度に基づいて前記第２の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、
これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、
この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項８】音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成

し、
これらの合成音声素片についてスペクトル整形を行い、
このスペクトル整形を行った後の各合成音声素片と前記第１の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、
前記距離尺度に基づいて前記第２の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、
これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成し、
この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。
【請求項９】前記代表音声素片として、音源信号と該音源信号を入力として合成音声信号を生成する合成フィルタの係数の組の情報を記憶することを特徴とする請求項１～８のいずれか１項に記載の音声合成方法。
【請求項１０】前記音源信号と前記合成フィルタの係数を量子化し、これら量子化した音源信号と合成フィルタの係数の組の情報を前記代表音声素片として記憶することを特徴とする請求項９に記載の音声合成方法。
【請求項１１】前記代表音声素片として、音源信号と該音源信号を入力として合成音声信号を生成する合成フィルタの係数の組の情報を記憶し、
かつ該代表音声素片の情報として記憶する音源信号および合成フィルタの係数のうちの少なくとも一方の数が音声合成単位の総数より少ないことを特徴とする請求項１～８のいずれか１項に記載の音声合成方法。
【請求項１２】前記代表音声素片として、音源信号と該音源信号を入力として合成音声信号を生成する合成フィルタの係数の組の情報を記憶し、
かつ該代表音声素片の情報として記憶する音源信号および合成フィルタの係数のうちの少なくとも一方の数が前記音素環境クラスタの総数より少ないことを特徴とする請求項２、３、６、７、８、９のいずれか１項に記載の音声合成方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、テキスト音声合成のための音声合成方法に係り、特に音韻記号列、ピッチおよび音韻継続時間長などの情報から音声信号を生成する音声合成方法に関する。
【０００２】
【従来の技術】任意の文章から人工的に音声信号を作り出すことをテキスト音声合成という。テキスト音声合成は、一般的に言語処理部、音韻処理部および音声合成部の３つの段階によって行われる。入力されたテキストは、まず言語処理部において形態素解析や構文解析などが行われ、次に音韻処理部においてアクセントやイントネーションの処理が行われて、音韻記号列・ピッチ・音韻継続時間長などの情報が出力される。最後に、音声信号合成部で音韻記号列・ピッチ・音韻継続時間長などの情報から音声信号を合成する。そこで、テキスト音声合成に用いる音声合成方法は、任意の音韻記号列を任意の韻律で音声合成することが可能な方法でなければならない。
【０００３】このような任意の音韻記号列を音声合成する音声合成装置の原理は、母音をＶ、子音をＣで表すと、ＣＶ、ＣＶＣ、ＶＣＶといった基本となる小さな単位の特徴パラメータ（これを代表音声素片という）を記憶し、これらを選択的に読み出した後、ピッチや継続時間長を制御して接続することにより、音声を合成するというものである。従って、記憶されている代表音声素片が合成音声の品質を大きく左右することになる。
【０００４】従来、これらの代表音声素片の作成はもっぱら人手に頼っており、音声信号の中から試行錯誤的に切り出してくる場合がほとんどであるため、膨大な労力を要していた。このような代表音声素片作成の作業を自動化し、音声合成に使用するのに適した代表音声素片を容易に生成する方法として、例えば音素環境クラスタリング（ＣＯＣ）と呼ばれる技術が特開昭６４－７８３００「音声合成方法」に開示されている。
【０００５】ＣＯＣの原理は、音素名や音素環境のラベルを多数の音声素片に付与し、そのラベルが付与された音声素片を音声素片間の距離尺度に基づいて音素環境に関する複数のクラスタに分類し、その各クラスタのセントロイドを代表音声素片とするものである。ここで、音素環境とは当該音声素片にとっての環境となる要因全ての組合せであり、その要因としては当該音声素片の音素名、先行音素、後続音素、後々続音素、ピッチ周期、パワー、ストレスの有無、アクセント核からの位置、息継ぎからの時間、発声速度、感情などが考えられる。実音声中の各音素は音素環境によって音韻が変化しているため、音素環境に関する複数のクラスタ毎に代表音声素片を記憶しておくことにより、音素環境の影響を考慮した自然な音声を合成することが可能となっている。
【０００６】
【発明が解決しようとする課題】上に述べたように、テキスト音声合成のための音声合成では、代表音声素片のピッチや継続時間長を指定された値に変更して合成する必要がある。このようなピッチや継続時間長の変更により、代表音声素片を切り出してきた音声信号の音質と比較して合成音声の音質がある程度劣化することになる。
【０００７】これに対して、上記のＣＯＣによるクラスタリングでは、音声素片間の距離尺度に基づいてクラスタリングを行っているにすぎないため、合成の際のピッチや継続時間の変更の効果が全く考慮されていないという問題がある。すなわち、ＣＯＣによるクラスタリングおよび各クラスタの代表音声素片は、実際にピッチや継続時間長を変更して合成された合成音声のレベルでは、

必ずしも適当なものになっているという保証はない。

【０００８】本発明は、このような問題点を解決すべくなされたものであり、テキスト音声合成による合成音声の音質を効果的に向上させることができる音声合成方法を提供することを目的とする。

【０００９】

【課題を解決するための手段】上記の課題を解決するため、本発明はピッチや継続時間長の変更の影響を考慮して、合成音声のレベルで自然音声に対する歪みが小さくなるような代表音声素片を生成し、その代表音声素片を用いて音声を合成することにより、自然音声に近い合成音声を生成するようにしたものである。

【００１０】すなわち、本発明に係る音声合成方法は、複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、これらの合成音声素片と第１の音声素片との間の距離尺度に基づいて第２の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする。

【００１１】ここで、第１および第２の音声素片は、ＣＶ，ＶＣＶ，ＣＶＣといった音声合成単位で音声信号中から切り出される素片であり、切り出された波形もしくはその波形から何らかの方法で抽出されたパラメータ系列などを表すものとする。これらのうち、第１の音声素片は合成音声の歪みを評価するために用いられ、また第２の音声素片は代表音声素片の候補として用いられる。合成音声素片は、第２の音声素片に対して少なくともピッチまたは継続時間長を変更して生成される合成音声波形またはパラメータ系列などを表す。

【００１２】合成音声素片と第１の音声素片との間の距離尺度によって、合成音声の歪みが表わされる。従って、この距離尺度つまり歪みがより小さくなる音声素片を第２の音声素片から選択して代表音声素片として記憶しておき、これらの代表音声素片から所定の代表音声素片を選択して接続すれば、自然音声に近い高品質の合成音声が生成される。

【００１３】本発明の第１の態様では、音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第１の音声素片との間の距離尺度に基づいて第２の音声素片から複数の代表音声素片を選択して記憶し、前記距離尺度に基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、複数の代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成する。

【００１４】ここで、音素環境とは前述した通り音声素片にとっての環境となる要因、例えば当該音声素片の音素名、先行音素、後続音素、後々続音素、ピッチ周期、パワー、ストレスの有無、アクセント核からの位置、息継ぎからの時間、発声速度、感情といった要素の組み合わせであり、音素環境クラスタとは言い換えれば音素環境の集合であり、例えば「当該素片の音韻が／ｋａ／、先行音韻が／ｉ／または／ｕ／、ピッチ周波数が２００Ｈｚ以下」というようなものを意味する。

【００１５】第１の態様のように、距離尺度つまり合成音声の歪みに基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続するようにすれば、例えば同一音素名の音声素片が複数の音素環境に存在する場合でも、実際の入力音素の音素環境が含まれる音素環境クラスタに対応する代表音声素片のみが選択されることにより、より自然な合成音声が得られる。

【００１６】本発明の第２の態様では、音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第１の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、前記距離尺度に基づいて第２の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成する。この第２の態様は、音声素片が一つの音素環境にのみ存在する場合に有効である。

【００１７】本発明の第３の態様では、音素環境がラベル付けされた複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第１の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、第１の音声素片と合成音声素片との間の距離尺度に基づいて第２の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成する。

【００１８】この第３の態様によっても、第１の態様と同様に、例えば同一音素名の音声素片が複数の音素環境に存在する場合、実際の入力音素の音素環境が含まれる音素環境クラスタに対応する代表音声素片のみが選択されることにより、より自然な合成音声が得られる。

【００１９】また、本発明に係る他の音声合成方法は、複数の第１の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第２の音声素片のピッチおよ

び継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、さらにこれらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と第１の音声素片との間の距離尺度に基づいて第２の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする。

【００２０】この場合、先に示した第１、第２および第３の態様においても、複数の合成音声素片を生成した後、スペクトル整形を行うようにする。ここで、スペクトル整形は「めりはり」のある明瞭な音声を合成するための処理であり、例えばホルマント強調やピッチ強調を行う適応ポストフィルタによるフィルタリングによって実現される。

【００２１】このように代表音声素片の接続によって合成される音声に対してスペクトル整形を行うと共に、合成音声素片に対しても同様のスペクトル整形を行うことによって、スペクトル整形後の最終的な合成音声のレベルで、自然音声に対する歪が小さくなるような代表音声素片を生成できるため、「めりはり」に優れたより明瞭な合成音声が得られる。

【００２２】本発明においては、代表音声素片として、音源信号と該音源信号を入力として合成音声信号を生成する合成フィルタの係数の組の情報を記憶するようにしてもよい。この場合、音源信号と合成フィルタの係数を量子化し、これら量子化した音源信号と合成フィルタの係数の組の情報を記憶するようにすれば、代表音声素片として記憶する音源信号と合成フィルタの係数の数を減少させることができるため、合成単位の学習に要する計算時間が短縮され、かつ実際の音声合成時に必要なメモリ量が低減される。

【００２３】さらに、代表音声素片の情報として記憶する音源信号および合成フィルタの係数のうちの少なくとも一方の数を音声合成単位の総数や、音素環境クラスタの総数より少なくすることも可能であり、このようにしても良好な合成音声を得ることができる。

【００２４】

【発明の実施の形態】以下、図面を参照して本発明の実施形態を説明する。

（第１の実施形態）図１は、本発明の第１の実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図である。この音声合成装置は、大きく分けて合成単位学習系１と規則合成系２からなる。実際にテキスト音声合成を行う場合に動作するのは規則合成系２であり、合成単位学習系１は事前に学習を行って代表音声素片を生成するものである。

【００２５】まず、合成単位学習系１について説明する。合成単位学習系１は、代表音声素片とこれに付随する音素環境クラスタを生成する代表音声素片生成部１１と代表音声素片記憶部１２および音素環境クラスタ記憶部１３により構成される。代表音声素片生成部１１には、第１の音声素片であるトレーニング音声素片１０１とこれにラベル付けされた音素環境１０２および第２の音声素片である入力音声素片１０３が入力される。

【００２６】代表音声素片生成部１１では、トレーニング音声素片１０１にラベル付けされた音素環境１０２に含まれるピッチ周期および継続時間長の情報に従って、入力音声素片１０３のピッチ周期および継続時間長を変更することで複数の合成音声素片が内部的に生成され、さらにこれらの合成音声素片とトレーニング音声素片１０１との距離尺度に従って、代表音声素片１０４と音素環境クラスタ１０５が生成される。音素環境クラスタ１０５は、トレーニング音声素片１０１を後述するように音素環境に関するクラスタに分類して生成される。

【００２７】代表音声素片１０４は代表音声素片記憶部１２に記憶され、音素環境クラスタ１０５は代表音声素片１０４と対応付けられて音素環境クラスタ記憶部１３に記憶される。代表音声素片生成部１１の処理については、後に詳細に説明する。

【００２８】次に、規則合成系２について説明する。規則合成系２は、代表音声素片記憶部１２と音素環境クラスタ記憶部１３と素片選択部１４および音声合成部１５により構成され、代表音声素片記憶部１２と音素環境クラスタ記憶部１３を合成単位学習系１と共有している。

【００２９】素片選択部１４には、入力音素の情報として、例えばテキスト音声合成のために入力テキストの形態素解析・構文解析後さらにアクセントやイントネーション処理を行って得られた韻律情報１１１と音韻記号列１１２が入力される。韻律情報１１１には、ピッチパターンおよび音韻継続時間長が含まれている。素片選択部１４では、これらの韻律情報１１１と音韻記号列１１２から入力音素の音素環境を内部的に生成する。

【００３０】そして、素片選択部１４は音素環境クラスタ記憶部１３より読み出された音素環境クラスタ１０６を参照して、入力音素の音素環境がどの音素環境クラスタに属するかを探索し、探索した音素環境クラスタに対応する代表音声素片選択情報１０７を代表音声素片記憶部１２へ出力する。

【００３１】音声合成部１５は、代表音声素片選択情報１０７に従って代表音声素片記憶部１２より選択的に読み出された代表音声素片１０８に対して、韻律情報１１１に従ってピッチ周期および音韻継続時間長を変更するとともに、素片の接続を行って合成音声信号１１３を出力する。ここで、ピッチおよび継続時間長を変更して素片を接続し音声を合成する方法としては、例えば残差駆動ＬＳＰ方法や波形編集方法など公知の技術を用いることができる。

【００３２】次に、本発明の特徴をなす代表音声素片生

成部１１の処理手順について具体的に説明する。図２のフローチャートは、代表音声素片生成部１１の第１の処理手順を示している。

【００３３】この第１の実施形態による代表音声素片生成処理では、まず準備段階として連続発声された多数の音声データに対して音韻毎にラベリングを行い、ＣＶ，ＶＣＶ，ＣＶＣなどの合成単位に従って、トレーニング音声素片T_i（ｉ＝１，２，３，…，N_T）を切り出す。また、各トレーニング音声素片T_iに対応する音素環境P_i（ｉ＝１，２，３，…，N_T）も抽出しておく。ただし、N_Tはトレーニング音声素片の個数を表す。音素環境P_iは、少なくともトレーニング音声素片T_iの音韻とそのピッチおよび継続時間長の情報を含むものとし、その他に必要に応じて前後の音素などの情報を含むものとする。

【００３４】次に、上述したトレーニング音声素片T_iの作成と同様の方法により、多数の入力音声素片S_j（ｉ＝１，２，３，…，N_S）を作成する。ただし、N_Sは入力音声素片の個数を表す。ここで、入力音声素片S_jとしてはトレーニング音声素片T_iと同じものを使用してもよいし（すなわちT_i＝S_i）、トレーニング音声素片T_iとは異なる音声素片を作成してもよい。いずれにしても、豊富な音韻環境を有する多数のトレーニング音声素片および入力音声素片が用意されていることが望ましい。

【００３５】このような準備段階を経た後、まず音声合成ステップＳ２１で、音素環境P_iに含まれるピッチおよび継続時間長に等しくなるように、入力音声素片S_jのピッチおよび継続時間長を変更して音声を合成することにより、合成音声素片G_{ij}を生成する。ここでのピッチおよび継続時間長の変更は、音声合成部１５におけるピッチおよび継続時間長の変更と同様の方法で行われるものとする。全ての音素環境P_i（ｉ＝１，２，３，…，N_T）に従って入力音声素片S_j（ｊ＝１，２，３，…，N_S）を用いて音声の合成を行うことにより、$N_T \times N_S$個の合成音声素片G_{ij}（ｉ＝１，２，３，…，N_T、ｊ＝１，２，３，…，N_S）を生成する。

【００３６】次に、歪み評価ステップＳ２２では、合成音声素片G_{ij}の歪みe_{ij}の評価を行う。この歪みe_{ij}の評価は、合成音声素片G_{ij}とトレーニング音声素片T_iとの間の距離尺度を求めることにより行う。距離尺度には、何らかのスペクトル距離を用いることができる。例えば、合成音声素片G_{ij}およびトレーニング音声素片T_iについて、ＦＦＴ（高速フーリエ変換）などを用いてパワースペクトルを求めて各パワースペクトル間の距離を評価する方法や、あるいは線形予測分析を行ってＬＰＣまたはＬＳＰパラメータなどを求めて各パラメータ間の距離を評価する方法などがある。その他にも、短時間フーリエ変換やウェーブレット変換などの変換係数を用いて評価する方法も用いることができる。また、各素片のパワーを正規化した上で歪みの評価を行う方法でもよい。

【００３７】次に、代表音声素片生成ステップＳ２３では、ステップＳ２２で得られた歪みe_{ij}に基づいて、入力音声素片S_jの中から指定された代表音声素片数Ｎの代表音声素片D_k（ｋ＝１，２，３，…，Ｎ）を選択する。

【００３８】代表音声素片選択法の一例を説明する。入力音声素片S_jの中から選択されたＮ個の音声素片の集合Ｕ＝｛u_k ｜ u_k ＝S_j（ｋ＝１，２，３，…，Ｎ）｝に対して、歪みの総和を表す評価関数E_{D1}（Ｕ）を次式（１）のように定義する。

【００３９】

【数１】

$$E_{D1}(U) = \sum_{i=1}^{N_T} \min(e_{ij1}, e_{ij2}, e_{ij3}, \cdots, e_{ijN}) \quad (1)$$

【００４０】ただし、ｍｉｎ（e_{ij1}，e_{ij2}，e_{ij3}，…，e_{ijN}）はe_{ij1}，e_{ij2}，e_{ij3}，…，e_{ijN}の中の最小値を表す関数である。集合Ｕの組合せはN_S！／｛Ｎ！（N_S－Ｎ）！｝通りあり、これらの音声素片の集合Ｕの中から評価関数E_{D1}（Ｕ）を最小にするＵを探索し、その要素u_kを代表音声素片D_kとする。

【００４１】最後に、音素環境クラスタ生成ステップＳ２４では、音素環境P_i、歪みe_{ij}および代表音声素片D_kより、音素環境に関する複数のクラスタ（音素環境クラスタ）C_k（ｋ＝１，２，３，…，Ｎ）を生成する。音素環境クラスタC_kは、例えば次式（２）で表されるクラスタリングの評価関数E_{C1}を最小化するクラスタを探索することによって得られる。

【００４２】

【数２】

$$E_{C1} = \sum_{k=1}^{N} \sum_{P_i \in C_k} e_{ijk} \quad (2)$$

【００４３】こうしてステップＳ２３およびＳ２４で生成された代表音声素片D_kおよび音素環境クラスタC_kは、図１の代表音声素片記憶部１２および音素環境クラスタ記憶部１３にそれぞれ記憶される。

【００４４】次に、図３のフローチャートを参照して代表音声素片生成部１１の第２の処理手順について説明す

る。この第２の処理手順による代表音声素片生成処理では、まず初期音素環境クラスタ生成ステップＳ３０において、何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、初期音素環境クラスタを生成する。音素環境のクラスタリングには、例えば音韻によるクラスタリングを行うことができる。

【００４５】そして、入力音声素片S_jおよびトレーニング音声素片T_iのうち音韻が一致する音声素片のみをそれぞれ用いて、図２のステップＳ２１，Ｓ２２，Ｓ２３，Ｓ２４と同様の合成音声素片生成ステップＳ３１、歪み評価ステップＳ３２、代表音声素片生成ステップＳ３３、音素環境クラスタ生成ステップＳ３４の処理を順次行い、全ての初期音素環境クラスタについて同様の操作を繰り返すことにより、全ての代表音声素片およびそれに対応する音素環境クラスタの生成を行う。こうして生成された代表音声素片および音素環境クラスタは、図１の代表音声素片記憶部１２および音素環境クラスタ記憶部１３にそれぞれ記憶される。

【００４６】ただし、各初期音素環境クラスタ当たりの代表音声素片数が１であれば、初期音素環境クラスタが代表音声素片の音素環境クラスタとなるため、音素環境クラスタ生成ステップＳ３４は不要となり、初期音素環境クラスタを音素環境クラスタ記憶部１３に記憶すればよい。

【００４７】次に、図４のフローチャートを参照して代表音声素片生成部１１の第３の処理手順を説明する。この第３の処理手順による代表音声素片生成処理では、図２に示した第１の処理手順と同様に音声合成ステップＳ４１および歪み評価ステップＳ４２を順次経た後、次の音素環境クラスタ生成ステップＳ４３において、音素環境P_iおよび歪みe_{ij}に基づいて音素環境に関するクラスタC_k（ｋ＝１，２，３，…，Ｎ）を生成する。音素環境クラスタC_kは、例えば次式（３）（４）で表わされるクラスタリングの評価関数E_{c2}を最小化するクラスタを探索することによって得られる。

【００４８】

【数３】

$$E_{c2}=\sum_{k=1}^{N}\min\{f(k,1),f(k,2),f(k,3),\cdots,f(k,N)\} \quad (3)$$

$$f(k,j)=\sum_{P_i\in C_k}e_{ij} \quad (4)$$

【００４９】次に、代表音声素片生成ステップＳ４４において、歪みe_{ij}に基づいて音素環境クラスタC_kのそれぞれに対応する代表音声素片D_kを入力音声素片S_jより選択する。この代表音声素片D_kは、入力音声素片S_jから例えば次式（５）で表される歪み評価関数$E_{D2}(j)$を最小化する音声素片を探索することによって得られる。

【００５０】

【数４】

$$E_{D2}(j)=\sum_{P_i\in C_k}e_{ij} \quad (5)$$

【００５１】なお、この第３の処理手順による代表音声素片生成処理を変形し、第２の処理手順と同様に、何らかの先見的な知識に基づいて予め生成した初期音素環境クラスタ毎に代表音声素片の生成および音素環境クラスタの生成を行うことも可能である。

【００５２】（第２の実施形態）次に、図５～図９を用いて本発明の第２の実施形態について説明する。図５は、第２の実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図である。図１と相対応する部分に同一の参照符号を付して相違点を中心に説明すると、本実施形態では音声合成部１５の後段に適応ポストフィルタ１６が追加されている点が第１の実施形態と異なり、これに加えて代表音声素片生成部１１における複数の合成音声素片の生成法も先の実施形態と異なっている。

【００５３】すなわち、代表音声素片生成部１１では第１の実施形態と同様に、トレーニング音声素片１０１にラベル付けされた音素環境１０２に含まれるピッチ周期および継続時間長の情報に従って、入力音声素片１０３のピッチ周期および継続時間長を変更することで複数の合成音声素片を内部的に生成した後、これらの合成音声素片に対して適応ポストフィルタによるフィルタリングを施してスペクトル整形を行う。そして、この適応ポストフィルタによりスペクトル整形を行った後の各合成音声素片とトレーニング音声素片１０１との距離尺度に従って、代表音声素片１０４と音素環境クラスタ１０５が生成される。音素環境クラスタ１０５は、先の実施形態と同様にトレーニング音声素片１０１を音素環境に関するクラスタに分類して生成される。

【００５４】なお、この代表音声素片生成部１１において音素環境１０２に含まれるピッチ周期および継続時間長の情報に従って入力音声素片１０３のピッチ周期および継続時間長を変更して生成される複数の合成音声素片に対してフィルタリングを施してスペクトル整形を行う適応ポストフィルタは、音声合成部１５の後段に配置される適応ポストフィルタ１６と同様の構成でよい。

【００５５】一方、音声合成部１５では第１の実施形態と同様に代表音声素片選択情報１０７に従って代表音声素片記憶部１２より選択的に読み出された代表音声素片１０８に対し、韻律情報１１１に従ってピッチ周期および音韻継続時間長を変更するとともに、素片の接続を行って合成音声信号１１３を生成するが、本実施形態ではこの合成音声信号１１３がさらに適応ポストフィルタ１６に入力され、ここで音質向上のためのスペクトル整形が行われた後、最終的な合成音声信号１１４が取り出される。

【００５６】図６に、適応ポストフィルタ１６の一構成例を示す。この適応ポストフィルタ１６は、ホルマント強調フィルタ２１とピッチ強調フィルタ２２を縦続配置して構成される。

【００５７】ホルマント強調フィルタ２１は、代表音声素片選択情報１０７に従って代表音声素片記憶部１２から選択的に読み出された代表音声素片１０８をＬＰＣ分析して得られるＬＰＣ係数に基づいて決定されるフィルタ係数に従って、音声合成部１５から入力される合成音声信号１１３をフィルタリングすることにより、スペクトルの山の部分を強調する処理を行う。一方、ピッチ強調フィルタ２２は、韻律情報１１１に含まれるピッチ周期に基づいて決定されるパラメータに従って、ホルマント強調フィルタ２１の出力をフィルタリングすることにより、音声信号のピッチを強調する処理を行う。なお、ホルマント強調フィルタ２１とピッチ強調フィルタ２２の配置順序は逆であってもよい。

【００５８】このような適応ポストフィルタ１６の適用によりスペクトルが整形され、「めりはり」のある明瞭な音声を再生可能な合成音声信号１１４が得られる。適応ポストフィルタ１６としては図６に示した構成のものに限られず、音声符号化や音声合成の分野で用いられる公知の技術に基づく種々の構成を採用することが可能である。

【００５９】このように本実施形態では、規則合成系２において音声合成部１５の後段に適応ポストフィルタ１６が配置される点を考慮して、合成単位学習系１においても代表音声素片生成部１１で音素環境１０２に含まれるピッチ周期および継続時間長の情報に従って入力音声素片１０３のピッチ周期および継続時間長を変更して生成される複数の合成音声素片に対し、同様に適応ポストフィルタによるフィルタリングを行っている。従って、適応ポストフィルタ１６を通した後の最終的な合成音声信号１１４と同様のレベルで、自然音声に対する歪みが小さくなるような代表音声素片を代表音声素片生成部１１において生成できるため、さらに自然音声に近い合成音声を生成することが可能となる。

【００６０】次に、図５における代表音声素片生成部１１の処理手順について具体的に説明する。図７、図８および図９のフローチャートは、図５における代表音声素片生成部１１の第１、第２および第３の処理手順を示している。図７、図８および図９では、先に説明した図２、図３および図４に示した処理手順における音声合成ステップＳ２１、Ｓ３１およびＳ４１の後に、ポストフィルタリングステップＳ２５、Ｓ３６およびＳ４５が追加されている。

【００６１】ポストフィルタリングステップＳ２５、Ｓ３６およびＳ４５では、前述した適応ポストフィルタによるフィルタリングを行う。すなわち、音声合成ステップＳ２１、Ｓ３１およびＳ４１で生成された合成音声素片$G_{i,j}$に対し、入力音声素片S_iをＬＰＣ分析して得られるＬＰＣ係数に基づいて決定されるフィルタ係数に従ってフィルタリングを行うことにより、スペクトルの山の部分を強調するホルマント強調を行う。また、このホルマント強調後の合成音声素片に対し、さらにトレーニング音声素片T_iのピッチ周期に基づいて決定されるパラメータに従ってフィルタリングを行うことにより、ピッチ強調を行う。

【００６２】このようにして、ポストフィルタリングステップＳ２５、Ｓ３６およびＳ４５において、スペクトル整形を行う。このポストフィルタリングステップＳ２５、Ｓ３６およびＳ４５は、前述したように規則合成系２において音声合成部１５の後段に設けられる適応ポストフィルタ１６により合成音声信号１１３のスペクトル整形を行って音質の向上を図るポストフィルタリングを行うことを前提に、合成単位の学習を可能とする処理であり、この処理を適応ポストフィルタ１６による処理と組み合わせることによって、最終的に「めりはり」のある明瞭な合成音声信号１１４が生成される。

【００６３】（第３の実施形態）次に、図１０～図１２を用いて本発明の第３の実施形態を説明する。図１０は、第２の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図である。

【００６４】本実施形態における合成単位学習系３０は、ＬＰＣ分析・逆フィルタ部３１、音源信号記憶部３２、ＬＰＣ係数記憶部３３、音源信号生成部３４、合成フィルタ３５、歪み計算部３６および最小歪探索部１７０からなる。この合成単位学習部３０には、トレーニング音声素片１０１と、これにラベル付けされた音素環境１０２および入力音声素片１０３が入力される。入力音声素片１０３は、ＬＰＣ分析・逆フィルタ部３１に入力され、ＬＰＣ分析が行われてＬＰＣ係数２０１と予測残差信号２０２が出力される。ＬＰＣ係数２０１はＬＰＣ係数記憶部３３に記憶され、また予測残差信号２０２は音源信号記憶部３２に記憶される。

【００６５】音源信号記憶部３２に記憶された予測残差信号は、最小歪み探索部３７からの指令に従って一つずつ読み出され、音源信号生成部３４においてトレーニング音声素片１０１の音素環境１０２に含まれるピッチパターンおよび音韻継続時間長の情報に従って、そのピッ

チ周期および継続時間長が変更されることにより、音源信号が生成される。このようにして生成された音源信号は、最小歪探索部３７からの指令に従ってＬＰＣ係数記憶部３３から読み出されたＬＰＣ係数をフィルタ係数とする合成フィルタ３５に入力され、合成音声素片が作成される。

【００６６】次に、歪計算部３６においてトレーニング音声素片１０１に対する合成音声素片の誤差つまり歪みが計算され、この歪みが最小歪探索部３７において評価される。最小歪探索部３７は、ＬＰＣ係数記憶部３３と音源信号記憶部３２にそれぞれ記憶されているＬＰＣ係数と予測残差信号の全ての組み合わせを出力するように指令を出して、それらの組み合わせに対応して合成フィルタ３５で合成音声素片を生成させる。そして、最小の歪みを与えるＬＰＣ係数と予測残差信号の組み合わせを見い出し、それを記憶する。

【００６７】次に、この合成単位学習系３０の動作を図１１のフローチャートを用いて説明する。まず、準備段階として連続発声された多数の音声データに音韻毎にラベリングを行い、ＣＶ，ＶＣＶ，ＣＶＣなどの合成単位に従って、トレーニング音声素片T_i（ｉ＝１，２，３，…，N_T）を切り出す。また、各トレーニング音声素片T_iに対応する音素環境P_i（ｉ＝１，２，３，…，N_T）を抽出しておく。ただし、N_Tはトレーニング音声素片T_iの個数を表す。音素環境は、少なくとも当該トレーニング音声素片の音韻とそのピッチパターンおよび継続時間長を含むものとし、その他に必要に応じて前後の音素などを含むものとする。

【００６８】次に、トレーニング音声素片を作成したのと同様の方法により、多数の入力音声素片S_i（ｉ＝１，２，３，…，N_S）を作成する。ただし、N_Sは入力音声素片S_iの個数を表す。ここで、入力音声素片S_iとトレーニング音声素片T_iの合成単位は一致させるものとする。例えば、あるＣＶ音節“ｋａ”の代表音声素片を作成する場合、多数の音声データから切り出された音節“ｋａ”から入力音声素片S_iとトレーニング音声素片T_iを設定する。なお、入力音声素片はトレーニング音声素片と同じものを使用してもよいし（すなわち、$T_i = S_i$）、異なる音声素片を作成してもよい。いずれにしても、豊富な音韻環境を有する多数のレーニング音声素片と入力音声素片が用意されていることが望ましい。

【００６９】このような準備段階の後、ＬＰＣ分析ステップＳ５１で入力音声素片S_i（ｉ＝１，２，３，…，N_S）のＬＰＣ分析を行い、ＬＰＣ係数a_i（ｉ＝１，２，３，…，N_S）を求めると共に、その係数に基づいた逆フィルタリングを行い、予測残差信号e_i（ｉ＝１，２，３，…，N_S）を求める。ただし、ａはｐをＬＰＣ分析の次数とすると、ｐ個の要素を持つベクトルである。

【００７０】次に、求められた予測残差信号を音源信号として、ＬＰＣ係数とともにステップＳ５２で保存する。次のＬＰＣ係数・音源信号組み合わせステップＳ５３では、保存されたＬＰＣ係数と音源信号の組み合わせを一組（a_i，e_j）作成する。

【００７１】この一組の組み合わせに対して、次の音声合成ステップＳ５４でP_kのピッチパターンおよび継続時間長に等しくなるようにe_jのピッチおよび継続時間長を変更して音源信号を生成した後、ＬＰＣ係数a_iを持つ合成フィルタでフィルタリング演算を行い、合成音声素片G_k（ｉ，ｊ）を生成する。

【００７２】このように、全てのP_k（ｋ＝１，２，３，…，N_T）に従って音声合成を行うことにより、N_T個の合成音声素片G_k（ｉ，ｊ），（ｋ＝１，２，３，…，N_T）を生成する。

【００７３】次の歪み評価ステップＳ５５では、合成音声素片G_k（ｉ，ｊ）とトレーニング音声素片T_kとの間の歪みE_k（ｉ，ｊ）とP_kに関する歪みの総和Ｅを次式（６）（７）により求める。

【００７４】

【数５】

$$E_k(i,j) = D(T_k, G_k(i,j)) \qquad (6)$$

$$E_k(i,j) = \sum_{k=1}^{N_T} E_k(i,j) \qquad (7)$$

【００７５】ここでＤは歪み関数であり、何らかのスペクトル距離を用いることができる。例えば、ＦＦＴなどを用いてパワースペクトルを求めて、その間の距離を求める方法や、あるいは線形予測分析を行ってＬＰＣまたはＬＳＰパラメータなどを求めて、パラメータ間の距離を評価する方法などがある。その他にも、短時間フーリエ変換やウェーブレット変換などの変換係数を用いて評価する方法が考えられる。また、各素片のパワーを正規化した上で、歪みの評価を行うことも考えられる。

【００７６】ステップＳ５３～Ｓ５５の処理をＬＰＣ係数と音源信号の全ての組み合わせ（a_i，e_j），（ｉ，ｊ＝１，２，３，…，N_S）について行い、歪み評価ステップＳ５５でＥ（ｉ，ｊ）の最小値を与えるｉ，ｊの組を探索する。

【００７７】次の代表音声素片生成ステップＳ５７では、Ｅ（ｉ，ｊ）の最小値を与えるｉ，ｊの組、また

は、対応する（ａｉ，ｅｊ）、または、（ａｉ，ｅｊ）から生成される波形を代表音声素片として保存する。ただし、この代表音声素片生成ステップは、代表音声素片を各合成単位毎に一組生成する場合の処理であり、Ｎ組生成したい場合は、次のようにする。まず、N_S ＊N_S 個の（ａｉ，ｅｊ）の組の中からＮ組選択した集合を式（８）と置き、歪みの総和を表す評価関数を式（９）のように定義する。

【００７８】

【数６】

$$U=\{(a_i, e_j)^m, m=1, 2, \cdots, N\} \quad (8)$$

$$ED(U)=\sum_{k=1}^{N_T} \min(E_k(i,j)^m, E_k(i,j)^2, \cdots, E_k(i,j)^N) \quad (9)$$

【００７９】ただし、ｍｉｎ（ ）は最小値を表す関数である。集合Ｕの組合せは、N_S ＊N_S C_S 通りあり、これらの集合Ｕの中から評価関数ＥＤ（Ｕ）を最小にするＵを探索し、その要素$(a_i, e_j)^k$を代表音声素片とする。

【００８０】次に、本実施形態における規則合成系について図１２を用いて説明する。本実施形態における規則合成系４０は、組み合せ記憶部４１、音源信号記憶部４２、ＬＰＣ係数記憶部４３、音源信号生成部４４および合成フィルタ４５からなる。規則合成部４０には、入力されたテキストの言語処理とそれに続く音韻処理の結果得られる韻律情報１１１と音韻記号列１１２が入力される。組み合せ記憶部４１、音源信号記憶部４２およびＬＰＣ係数記憶部４３には、図１０の合成単位学習部３０で求められたＬＰＣ係数と音源信号の組み合わせ情報（ｉ，ｊ）と、音源信号ｅｊ，ＬＰＣ係数ａｉが予め記憶されている。

【００８１】組み合せ記憶部４１は、音韻記号列１１２を入力し、これに対応する合成単位（例えばＣＶ音節）を与えるＬＰＣ係数と音源信号の組合せ情報を出力する。音源信号記憶部４２に記憶された音源信号は、組み合せ記憶部４１からの指令に従って読み出され、音源信号生成部４４において入力された韻律情報１１１に含まれるピッチパターンおよび音韻継続時間長の情報に従って、そのピッチ周期および継続時間長が変更されると共に音源信号の接続が行われる。

【００８２】こうして生成された音源信号は、組み合せ記憶部４１の指令に従ってＬＰＣ係数記憶部４３から読み出された係数をフィルタ係数とする合成フィルタ４５に入力され、フィルタ係数の補間とフィルタリング演算が行われることにより、合成音声信号１１３が作成される。

【００８３】（第４の実施形態）次に、図１３および図１４を用いて本発明の第４の実施形態を説明する。図１３は本実施形態における合成単位学習系の概略構成を示す図であり、第３の実施形態の図１０に示した合成単位学習系３０にクラスタリング部３８を付加した構成となっている。本実施形態では、クラスタリング部３８において何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、各クラスタに対して代表音声素片を生成する点が第３の実施形態と異なる。クラスタリングとしては、例えば当該素片のピッチによるクラスタリングが考えられる。この場合、トレーニング音声素片１０１をピッチに基づいてクラスタリングし、各クラスタのトレーニング音声素片に対して第３の実施形態で述べた代表音声素片の生成を行う。

【００８４】図１４は、本実施形態における規則合成系の概略構成を示す図であり、第３の実施形態の図１２に示した規則合成系４０にクラスタリング部４８を付加した構成となっている。韻律情報１１１をトレーニング音声素片と同様にピッチクラスタリングし、合成単位学習系３０で求められた各クラスタの代表音声素片に対応する音源信号及びＬＰＣ係数を用いて音声を合成する。

【００８５】（第５の実施形態）次に、図１５～図１７を用いて本発明の第５の実施形態を説明する。図１５は、本実施形態における合成単位学習系を示すブロック図であり、クラスタをトレーニング音声素片との歪み尺度に基づいて自動的に生成する場合の構成例を示している。本実施形態は、図１０に示した合成単位学習系３０に音素環境クラスタ生成部５１とクラスタ記憶部５２が追加された構成となっている。

【００８６】本実施形態における合成単位学習系の第１の処理手順を図１６に示すフローチャートを用いて説明する。この処理手順は第３の実施形態の処理手順を示した図１１に新たに音素環境クラスタ生成ステップＳ５８でが追加されている。このステップＳ５８では、音素環境P_kと歪みE_k（ｉ，ｊ）および代表音声素片D_oより、音素環境に関するクラスタC_o（ｍ＝１，２，３，…，Ｎ）を生成する。音素環境クラスタC_oは、例えば次式（１０）で表されるクラスタリングの評価関数E_cを最小化するクラスタを探索することによって得られる。

【００８７】

【数７】

19　　20

$$E_{cm}=\sum_{m=1}^{N}\sum_{P_k\in C_m}E_k(i,j) \qquad (10)$$

【００８８】図１７は、図１５の合成単位学習系の第２の処理手順を示すフローチャートである。この処理では、初期音素環境クラスタ生成ステップＳ５０で何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、初期音素環境クラスタを生成する。コラスタリングとしては、例えば当該素片の音韻によるクラスタリングが考えられる。この場合、当該素片の音韻が一致する音声素片およびトレーニング音声素片だけを用いて第３の実施形態で述べた代表音声素片の生成および音素環境クラスタの生成を行い、全ての初期音素環境クラスタについて同様の操作を繰り返すことによって、全ての代表音声素片および対応する音素環境クラスタの生成を行う。

【００８９】ただし、各初期クラスタ当たりの代表音声素片数が１であれば、初期音素環境クラスタが代表音声素片の音素環境クラスタとなるため、音素環境クラスタ生成ステップＳ５８は不要となり、初期音素環境クラスタを図１５のクラスタ記憶部５２に記憶すればよい。

【００９０】本実施形態における規則合成系は、図１４に示した第４の実施形態における規則合計系４０と同様に構成される。この場合、クラスタリング部４８は図１５のクラスタ記憶部５２に蓄積された情報に基づいて処理を行う。

【００９１】（第６の実施形態）図１８に、本発明の第６の実施形態における合成単位学習系の構成を示す。本実施形態における合成単位学習系は、図１０に示した合成単位学習系３０にバッファ６１，６２および量子化テーブル作成部６３，６４が追加された構成となっている。

【００９２】本実施形態において、入力音声素片１０３はＬＰＣ分析・逆フィルタ部３１に入力され、ここでＬＰＣ分析により生成されたＬＰＣ係数２０１と予測残差信号２０２が一旦バッファ６１，６２にそれぞれ蓄えられた後、量子化テーブル作成部６３，６４でそれぞれ量子化され、量子化されたＬＰＣ係数と予測残差信号がＬＰＣ係数記憶部３３および音源信号記憶部３４にそれぞれ記憶される。

【００９３】図１９は、図１８の合成単位学習系の処理手順を示すフローチャートであり、図１１のフローチャートに示した処理手順と異なるところは、ＬＰＣ分析ステップＳ５１の後に量子化ステップＳ６０が追加されたことである。この量子化ステップＳ６０では、ＬＰＣ分析ステップＳ５１で求められたＬＰＣ係数a_i（ｉ＝１，２，３，…，N_s）と予測残差信号e_i（１＝１，２，３，…，N_s）を一旦バッファに蓄積した後、ＬＢＧアルゴリズムなどの公知の技術を用いて量子化テーブルを作成し、ＬＰＣ係数と予測残差信号を量子化する。このとき、量子化テーブルのサイズ、すなわち量子化の代表ベクトルの数はN_s未満とする。そして、量子化されたＬＰＣ係数と予測残差信号が次のステップＳ５２で保存される。その後の処理は、図１１の場合と同一である。

【００９４】（第７の実施形態）図２０は、本発明の第７の実施形態における合成単位学習系を示すブロック図であり、クラスタをトレーニング音声素片との歪み尺度に基づいて自動的に生成する場合の構成例を示している。クラスタの生成は、第５の実施形態と同様に行うことができる。すなわち、本実施形態における合成単位学習系は、図１５に示した第５の実施形態と図１８に示した第６の実施形態とを組み合わせた構成となっている。

【００９５】（第８の実施形態）図２１は、本発明の第８の実施形態における合成単位学習系であり、ＬＰＣ分析部３１ａと逆フィルタ３１ｂを分離して、バッファ６１および量子化テーブル作成部６３を経て量子化されたＬＰＣ係数を用いて逆フイルタリングを行って予測残差信号を計算する場合の構成例を示している。このようにすることにより、ＬＰＣ係数の量子化歪みによる合成音声の音質劣化を低減する代表音声素片を生成することが可能になる。

【００９６】（第９の実施形態）図２２は、本発明の第９の実施形態における合成単位学習系であり、第８の実施形態と同様に、量子化されたＬＰＣ係数を用いて逆フイルタリングし、予測残差信号を計算する場合の他の構成例を示している。ただし、本実施形態では逆フィルタ３１ｂで逆フィルタリングされた予測残差信号がバッファ６２および量子化テーブル６４を経て量子化された後、音源信号記憶部３２に入力される点が第８の実施形態と異なっている。

【００９７】第６～第９の実施形態において、量子化テーブル作成部６３，６４で作成される量子化テーブルのサイズ、すなわち量子化の代表スペクトルの数は、クラスタ数または合成単位の総数（例えば、ＣＶ，ＶＣ音節の総数）より少なく選ぶことができる。このようにＬＰＣ係数と予測残差信号を量子化することによって、代表音声素片として記憶されるＬＰＣ係数と音源信号の数を減少させることができるため、合成単位の学習に要する計算時間を短縮することができると共に、規則合成系で用いるメモリ量を低減することができる。

【００９８】しかも、ＬＰＣ係数と音源信号の組み合せ（a_i，e_i）で音声合成を行うので、ＬＰＣ係数と音源信号数のどちらかの代表音声素片数がクラスタ数や合成単位の総数（例えば、ＣＶ，ＶＣ音節の総数）より少ない場合でも、良好な合成音声を得ることができる。

【００９９】また、第６～第９の実施形態において、ト

レーニング音声素片と合成音声素片との歪み尺度として合成素片間の接続歪みを考慮することにより、より滑らかな合成音を得ることもできる。

【０１００】さらに、合成単位の学習および規則合成において、第２の実施形態で説明したと同様の適応ポストフィルタを合成フィルタと合わせて用いることもでき、これにより合成音声のスペクトルが整形され、「めりはり」のある明瞭な合成音声を得ることができる。

【０１０１】

【発明の効果】以上説明したように、本発明の音声合成方法によれば、入力音声素片に対してピッチおよび継続時間長の少なくとも一方の変更を行って生成される合成音声のレベルで自然音声に対する歪みを評価し、それに基づいて入力音声素片から選択した音声素片を代表音声素片とするため、音声合成装置の特性をも考慮した代表音声素片の生成が可能であり、この代表音声素片を接続して音声合成を行うことによって、自然音声に近い高品質の合成音声を生成することができる。

【０１０２】また、本発明ではさらに代表音声素片の接続によって合成される音声に対してスペクトル整形を行うと共に、合成音声素片に対しても同様のスペクトル整形を行うことにより、スペクトル整形後の最終的な合成音声信号のレベルで、自然音声に対する歪が小さくなるような代表音声素片を生成できるため、「めりはり」のあるより明瞭な合成音声を生成することができる。

【０１０３】また、各代表音声素片を音素環境に基づく素片選択規則に従って選択して接続することにより、合成音声は滑らかで自然性の高いものとなる。さらに、代表音声素片として音源信号（例えば予測残差信号）音源信号を入力として合成音声信号を生成する合成フィルタの係数（例えばＬＰＣ係数）の組の情報を記憶する場合、これらを量子化することによって、代表音声素片として記憶する音源信号と合成フィルタの係数の数を減少させることができるため、合成単位の学習に要する計算時間を短縮することができると共に、規則合成系で用いるメモリ量を低減することができる。

【０１０４】しかも、代表音声素片の情報として記憶する音源信号および合成フィルタの係数のうちの少なくとも一方の数を音声合成単位の総数（例えば、ＣＶ，ＶＣ音節の総数）や音素環境クラスタ数より少ない場合でも、良好な合成音声を得ることができる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態に係る音声合成装置の構成を示すブロック図

【図２】図１中の代表音声素片生成部での第１の処理手順を示すフローチャート

【図３】図１の中の代表音声素片生成部での第２の処理手順を示すフローチャート

【図４】図１中の代表音声素片生成部での第３の処理手順を示すフローチャート

【図５】本発明の第２の実施形態に係る音声合成装置の構成を示すブロック図

【図６】図５中の適応ポストフィルタの構成例を示すブロック図

【図７】図５中の代表音声素片生成部での第１の処理手順を示すフローチャート

【図８】図５の中の代表音声素片生成部での第２の処理手順を示すフローチャート

【図９】図５中の代表音声素片生成部での第３の処理手順を示すフローチャート

【図１０】本発明の第３の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図１１】図１０の合成単位学習系の処理手順を示すフローチャート

【図１２】本発明の第３の実施形態に係る音声合成装置における規則合成系の構成を示すブロック図

【図１３】本発明の第４の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図１４】本発明の第４の実施形態に係る音声合成装置における規則合成系の構成を示すブロック図

【図１５】本発明の第５の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図１６】図１５の合成単位学習系の第１の処理手順を示すフローチャート

【図１７】図１５の合成単位学習系の第２の処理手順を示すフローチャート

【図１８】本発明の第６の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図１９】図１８の合成単位学習系の処理手順を示すフローチャート

【図２０】本発明の第７の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図２１】本発明の第８の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図２２】本発明の第９の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【符号の説明】

１…合成単位学習系
２…規則合成系
１１…代表音声素片生成部
１２…音素環境クラスタ記憶部
１３…代表音声素片記憶部
１４…素片選択部
１５…音声合成部
１６…適応ポストフィルタ
２１…ホルマント強調フィルタ
２２…ピッチ強調フィルタ
１０１…トレーニング音声素片（第１の音声素片）
１０２…トレーニング音声素片にラベル付けされた音素環境

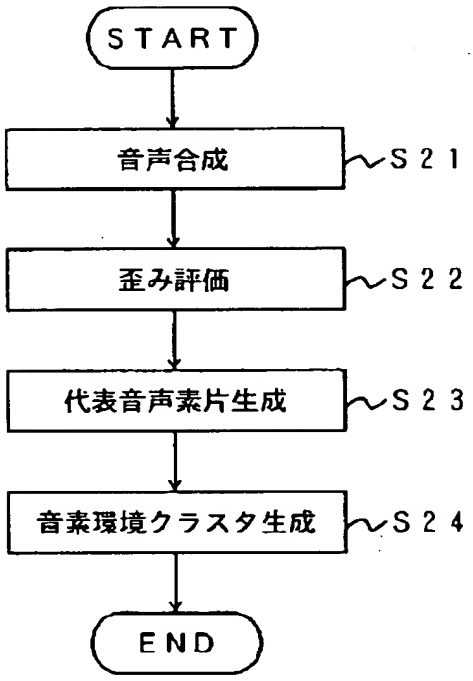
23

１０３…入力音声素片（第２の音声素片）
１０４…代表音声素片
１０５…音素環境クラスタ
１０６…音素環境クラスタ
１０７…代表音声素片選択情報
１０８…代表音声素片
１１１…韻律情報
１１２…音韻記号列
１１３…合成音声信号
１１４…合成音声信号
３０…合成単位学習系
３１…ＬＰＣ分析・逆フィルタ
３１ａ…ＬＰＣ分析部
３１ｂ…逆フィルタ
３２…音源信号記憶部
３３…ＬＰＣ係数記憶部

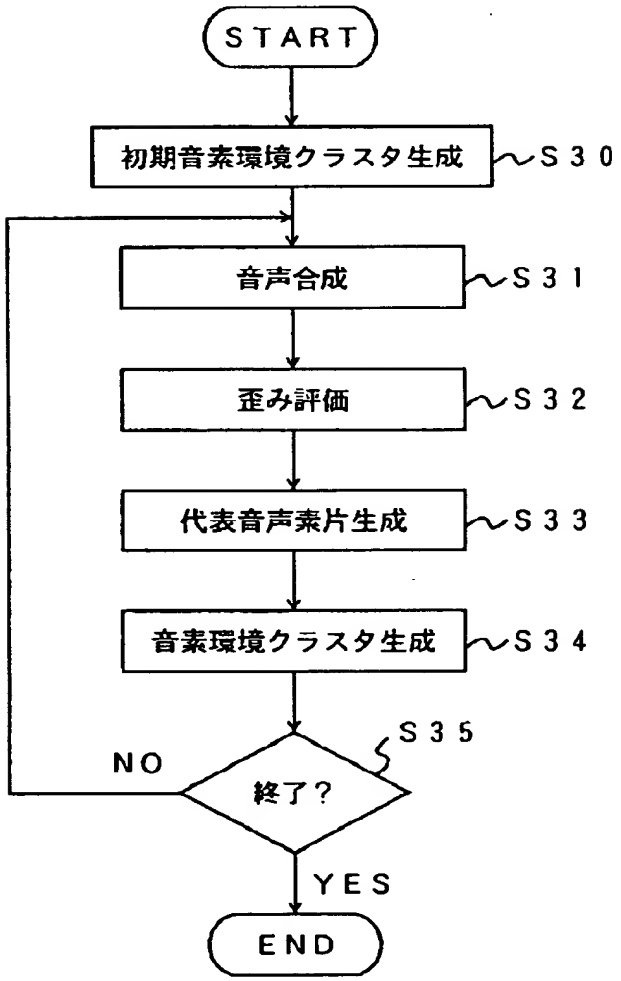
24

３４…音源信号生成部
３５…合成フィルタ
３６…歪計算部
３７…最小歪探索部
３８…クラスタリング部
４０…規則合成系
４１…組み合わせ記憶部
４２…音源信号記憶部
４３…ＬＰＣ係数記憶部
４４…音源信号生成部
４５…合成フィルタ
４８…クラスタリング部
５１…音素環境クラスタ生成部
５２…クラスタ記憶部
６１，６２…バッファ
６３，６４…量子化テーブル作成部

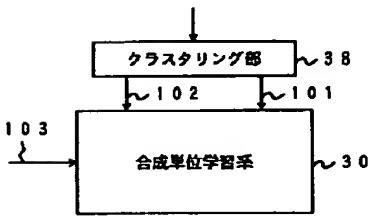
【図２】



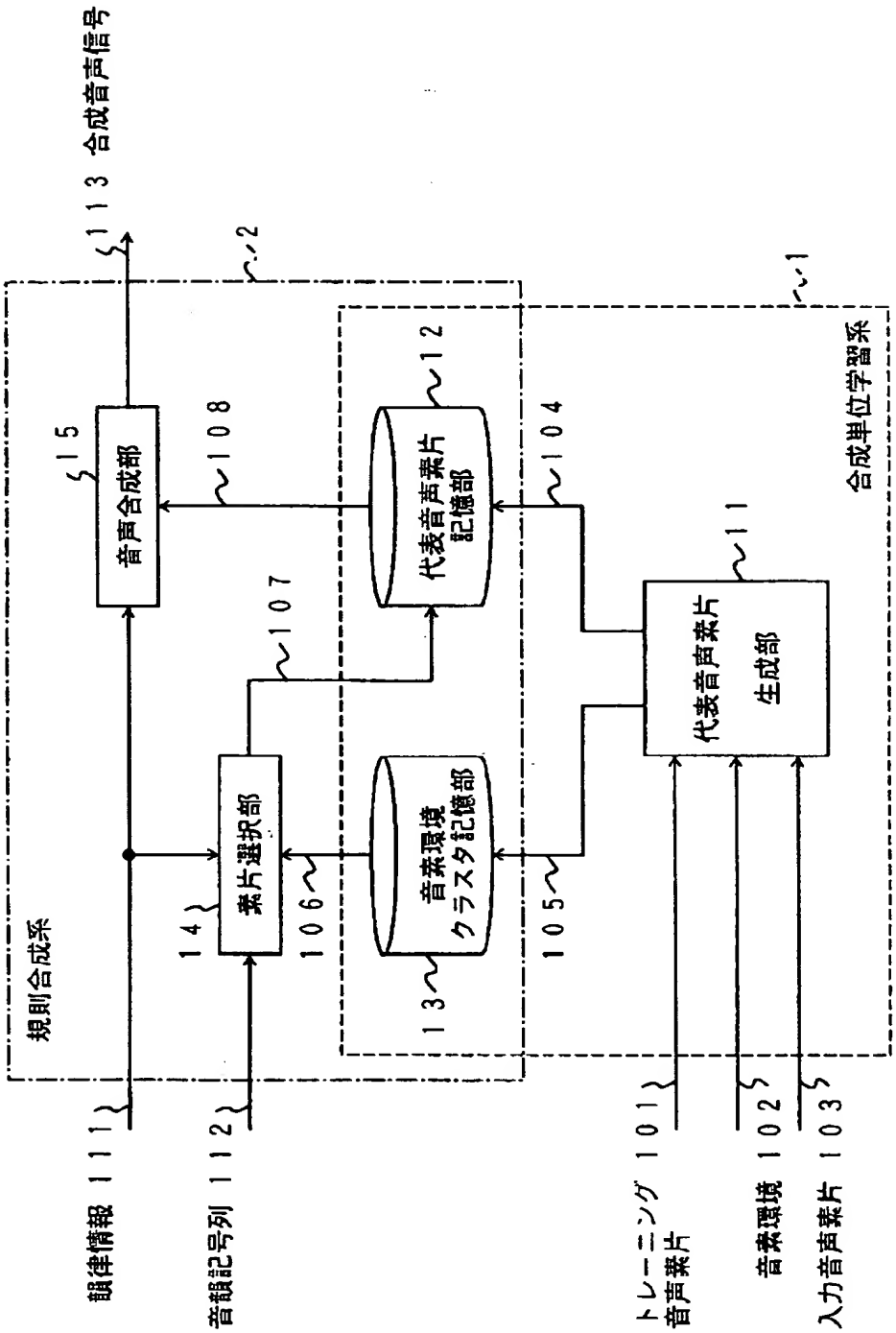
【図３】



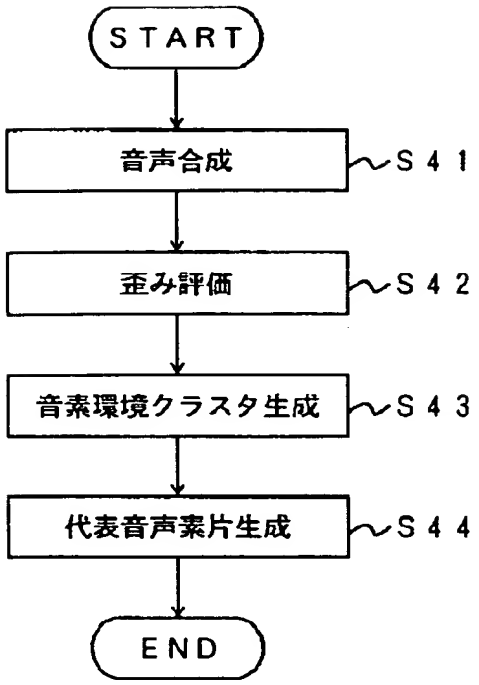
【図１３】



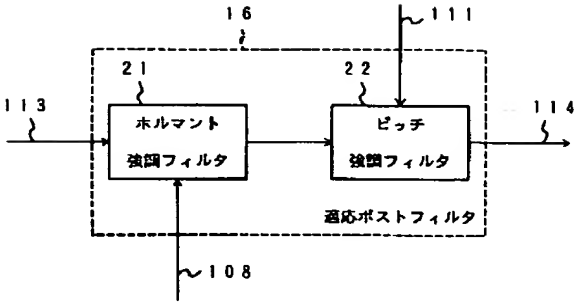
【図１】


113 合成音声信号
2
1
規則合成系
合成単位学習系
15
音声合成部
108
12
代表音声素片記憶部
104
107
11
代表音声素片生成部
14
素片選択部
106
音素環境クラスタ記憶部
105
13
韻律情報 111
音韻記号列 112
トレーニング音声素片 101
音素環境 102
入力音声素片 103

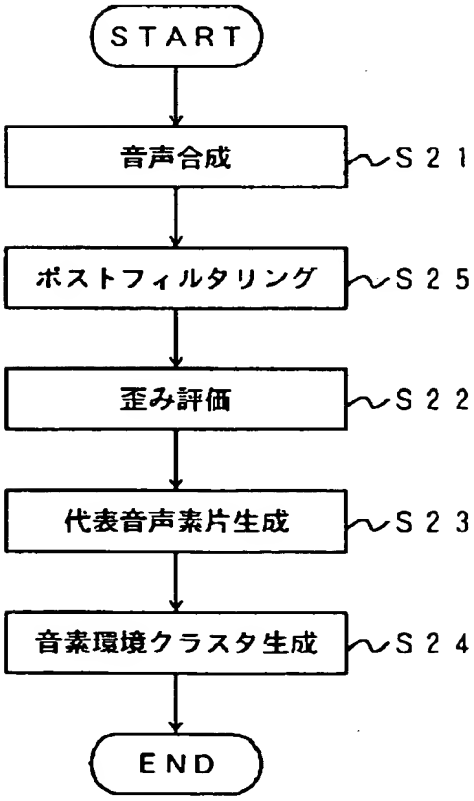
【図４】


START
音声合成
S41
歪み評価
S42
音素環境クラスタ生成
S43
代表音声素片生成
S44
END

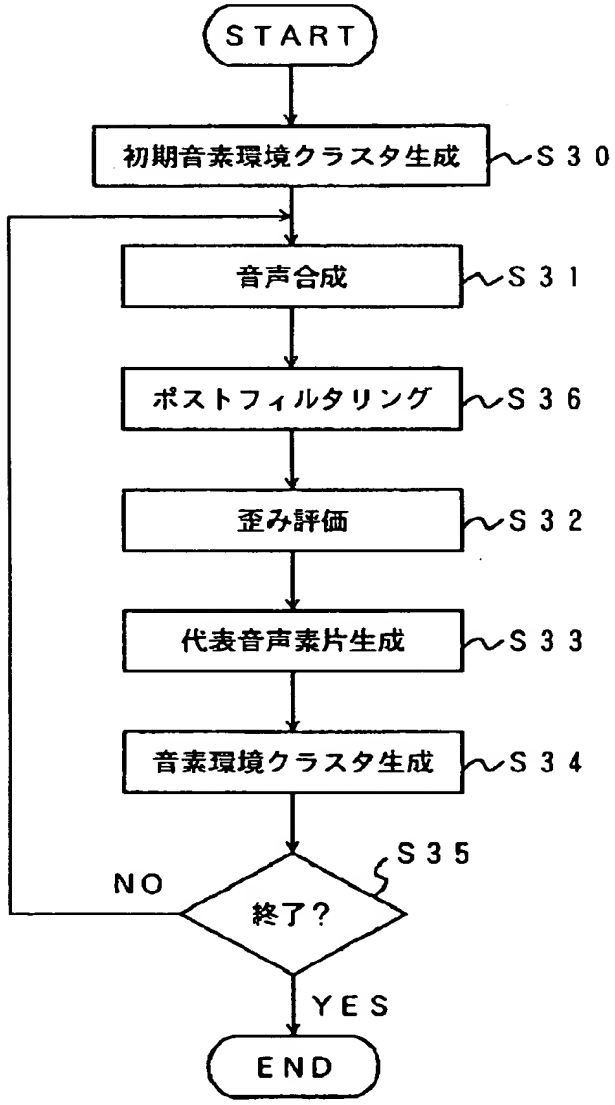
【図６】


16
111
21
22
113
ホルマント
強調フィルタ
ピッチ
強調フィルタ
114
適応ポストフィルタ
108

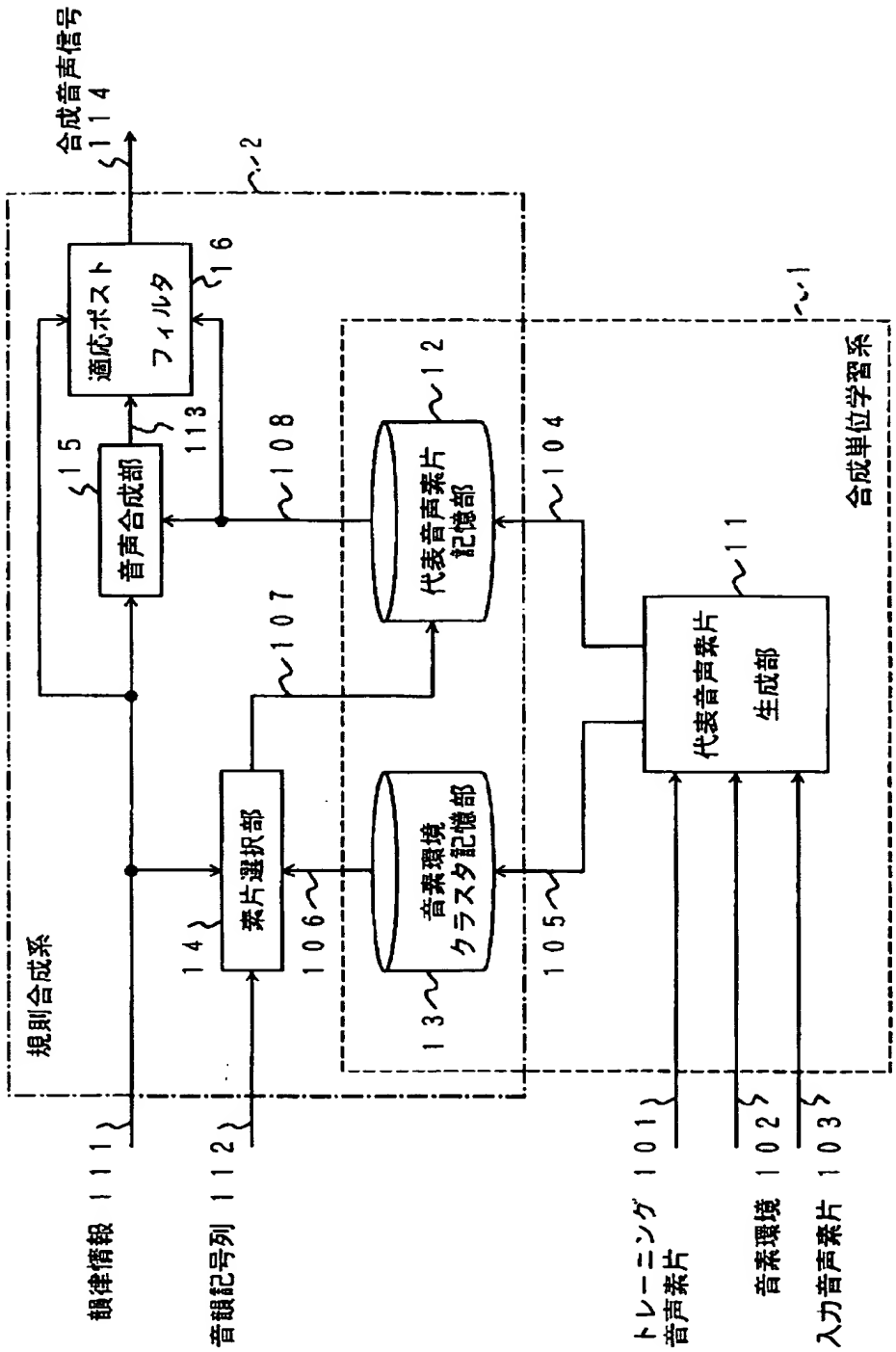
【図７】


START
音声合成
S21
ポストフィルタリング
S25
歪み評価
S22
代表音声素片生成
S23
音素環境クラスタ生成
S24
END

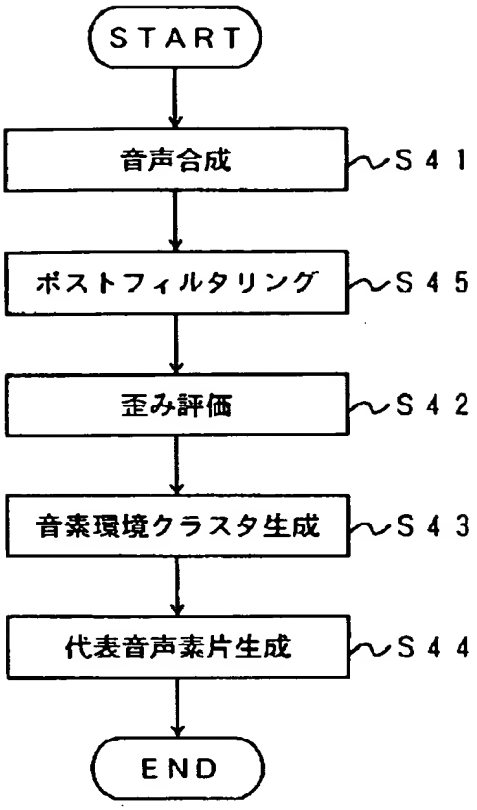
【図８】


START
初期音素環境クラスタ生成
S30
音声合成
S31
ポストフィルタリング
S36
歪み評価
S32
代表音声素片生成
S33
音素環境クラスタ生成
S34
S35
NO
終了？
YES
END

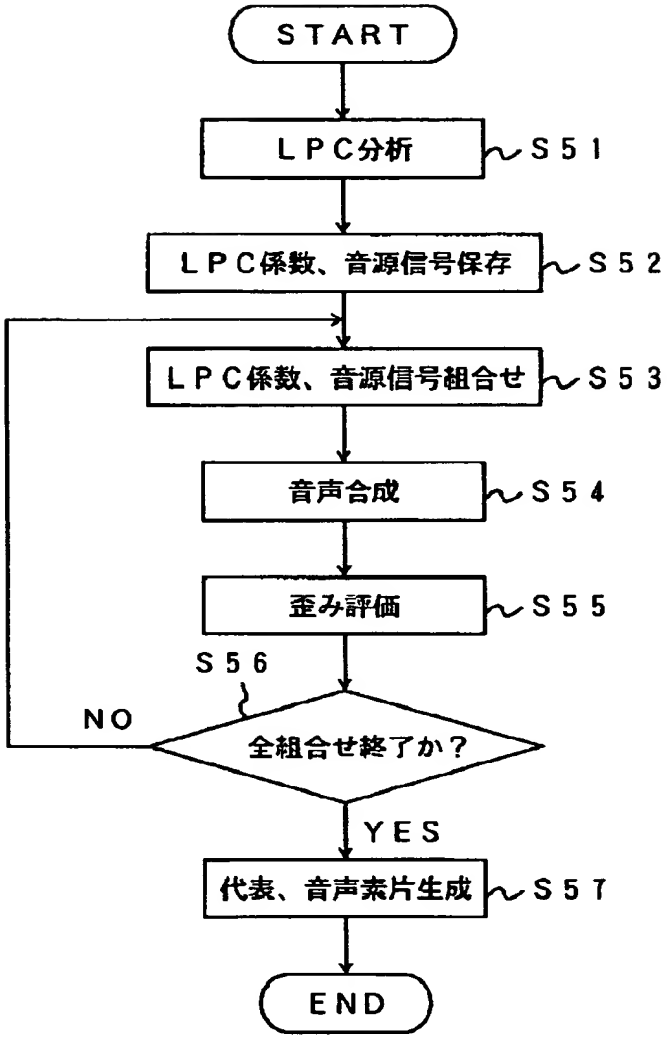
【図５】


合成音声信号
114
適応ポスト
フィルタ
16
音声合成部
15
113
108
代表音声素片
記憶部
12
104
107
素片選択部
14
106
音素環境
クラスタ記憶部
13
105
代表音声素片
生成部
11
規則合成系
合成単位学習系
韻律情報 111
音韻記号列 112
トレーニング
音声素片 101
音素環境 102
入力音声素片 103

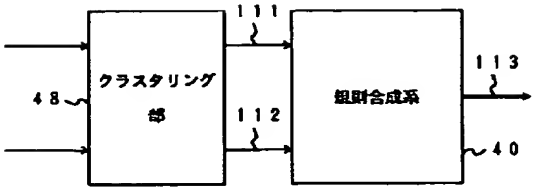
【図９】


START
音声合成
S41
ポストフィルタリング
S45
歪み評価
S42
音素環境クラスタ生成
S43
代表音声素片生成
S44
END

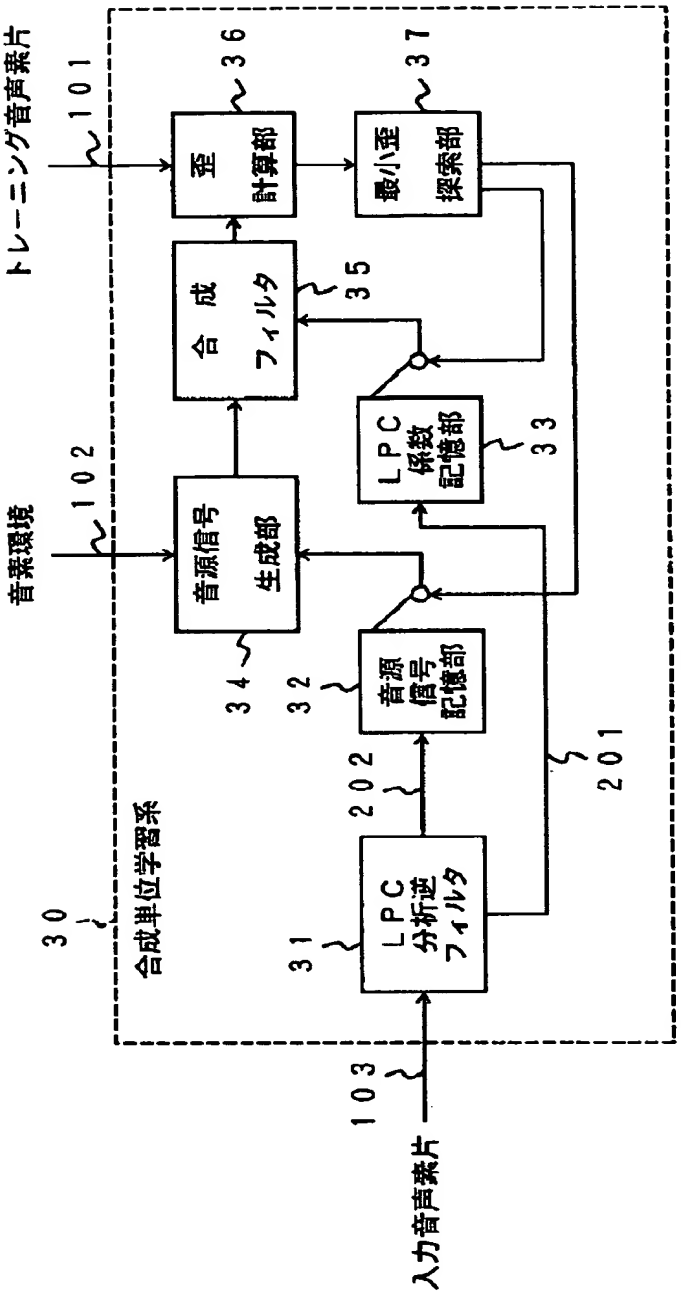
【図１１】


START
LPC分析
S51
LPC係数、音源信号保存
S52
LPC係数、音源信号組合せ
S53
音声合成
S54
歪み評価
S55
S56
NO
全組合せ終了か？
YES
代表、音声素片生成
S57
END

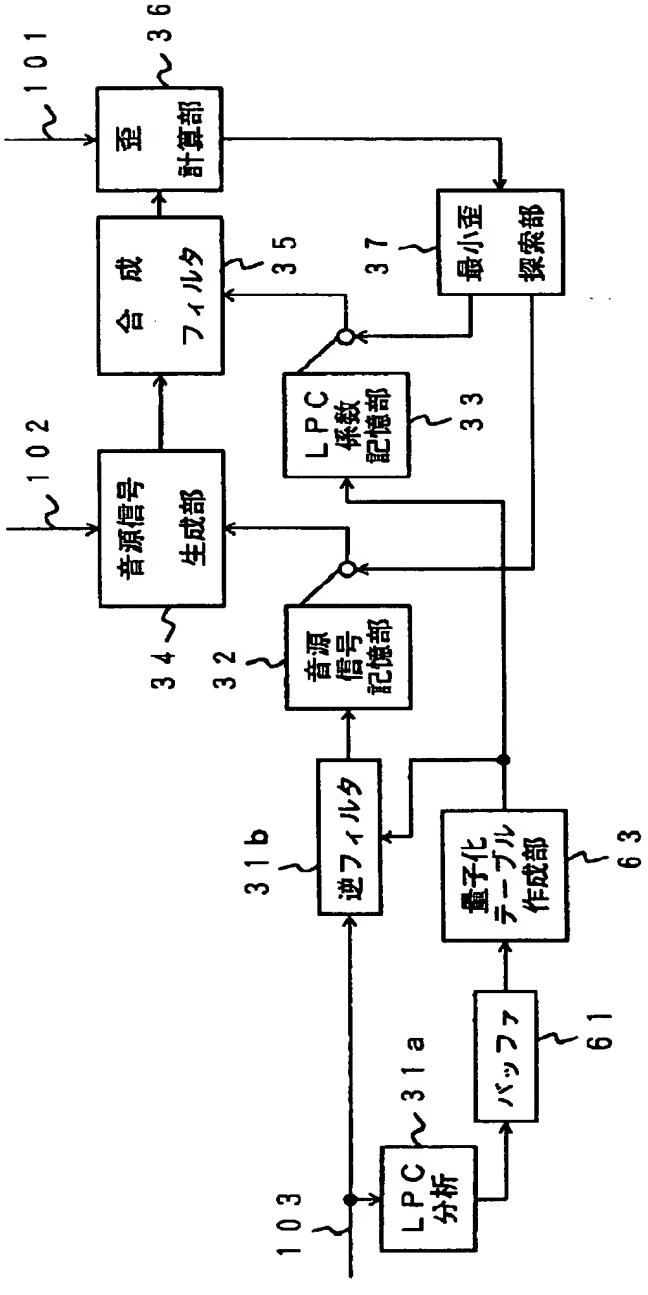
【図１４】


48
クラスタリング部
111
112
組則合成系
113
40

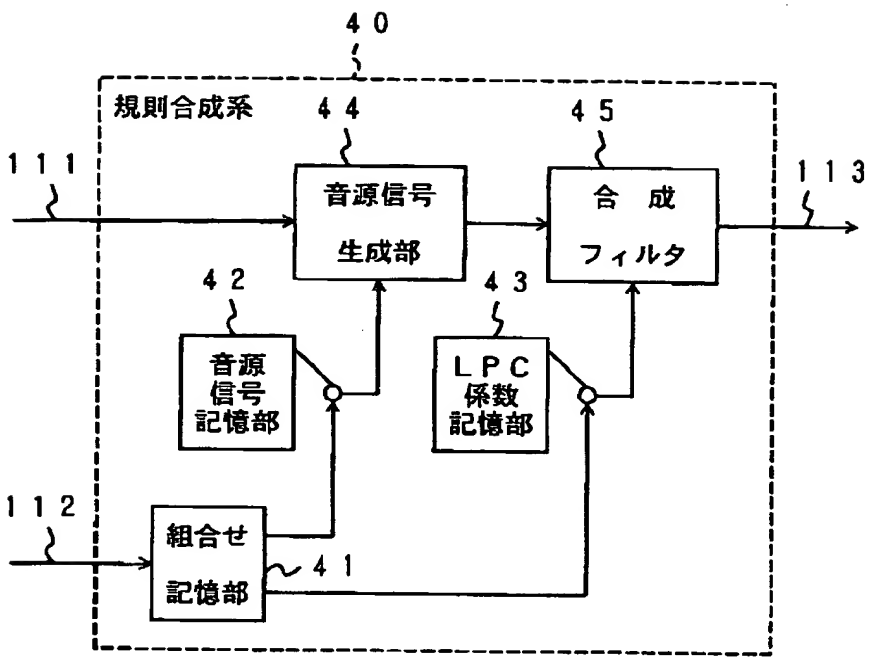
【図１０】


トレーニング音声素片
101
音素環境
102
合成単位学習系
30
歪 計算部
36
最小歪 探索部
37
合成 フィルタ
35
LPC 係数 記憶部
33
音源信号 生成部
34
音源 信号 記憶部
32
LPC 分析逆 フィルタ
31
202
201
103
入力音声素片

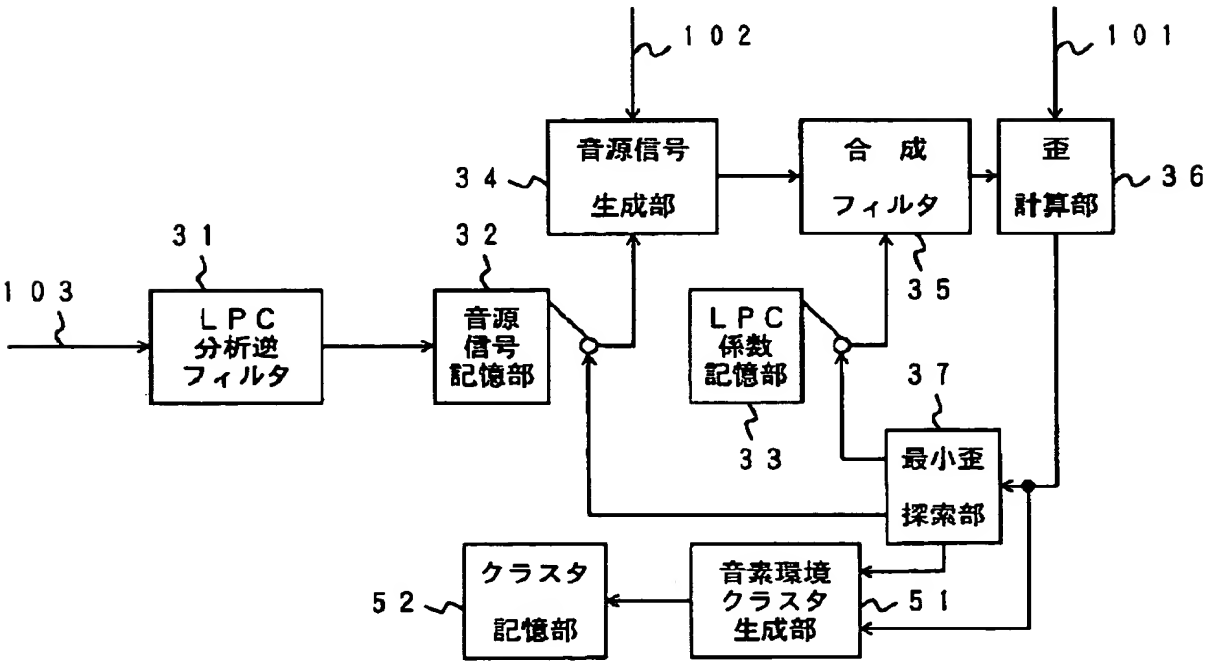
【図２１】


101
歪 計算部
36
合成 フィルタ
35
最小歪 探索部
37
LPC 係数 記憶部
33
102
音源信号 生成部
34
音源 信号 記憶部
32
31b
逆フィルタ
量子化 テーブル 作成部
63
バッファ
61
LPC 分析
31a
103

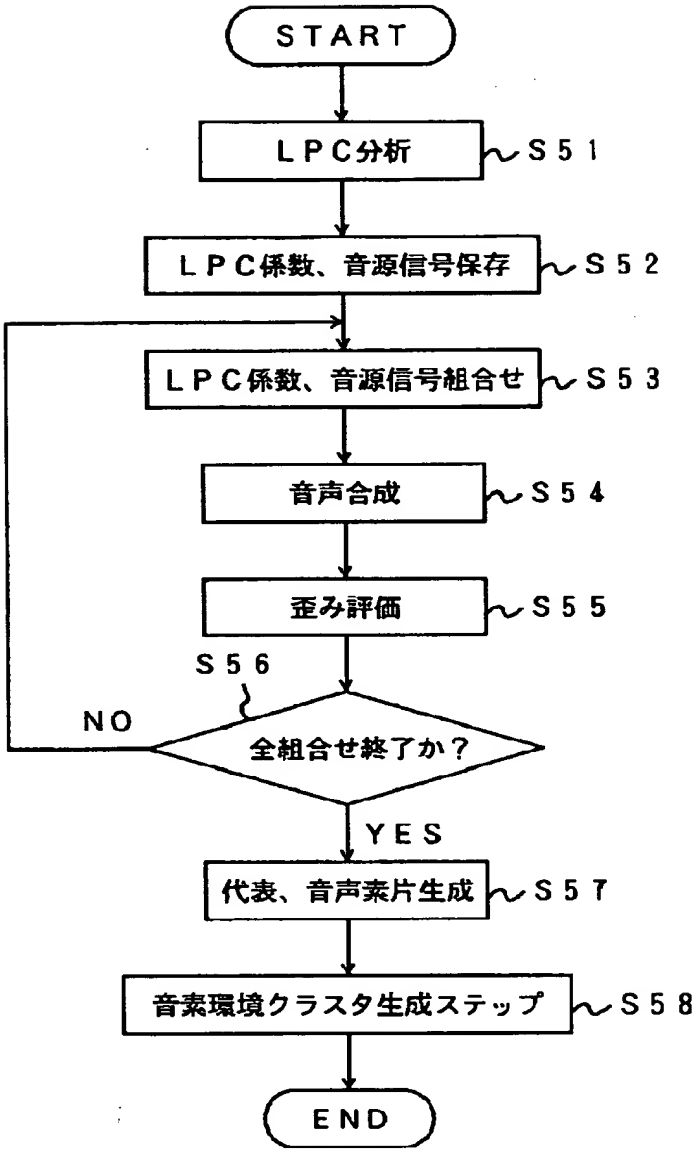
【図１２】


４０
規則合成系
４４
４５
１１１
音源信号
生成部
合　成
フィルタ
１１３
４２
４３
音源
信号
記憶部
ＬＰＣ
係数
記憶部
１１２
組合せ
記憶部
４１

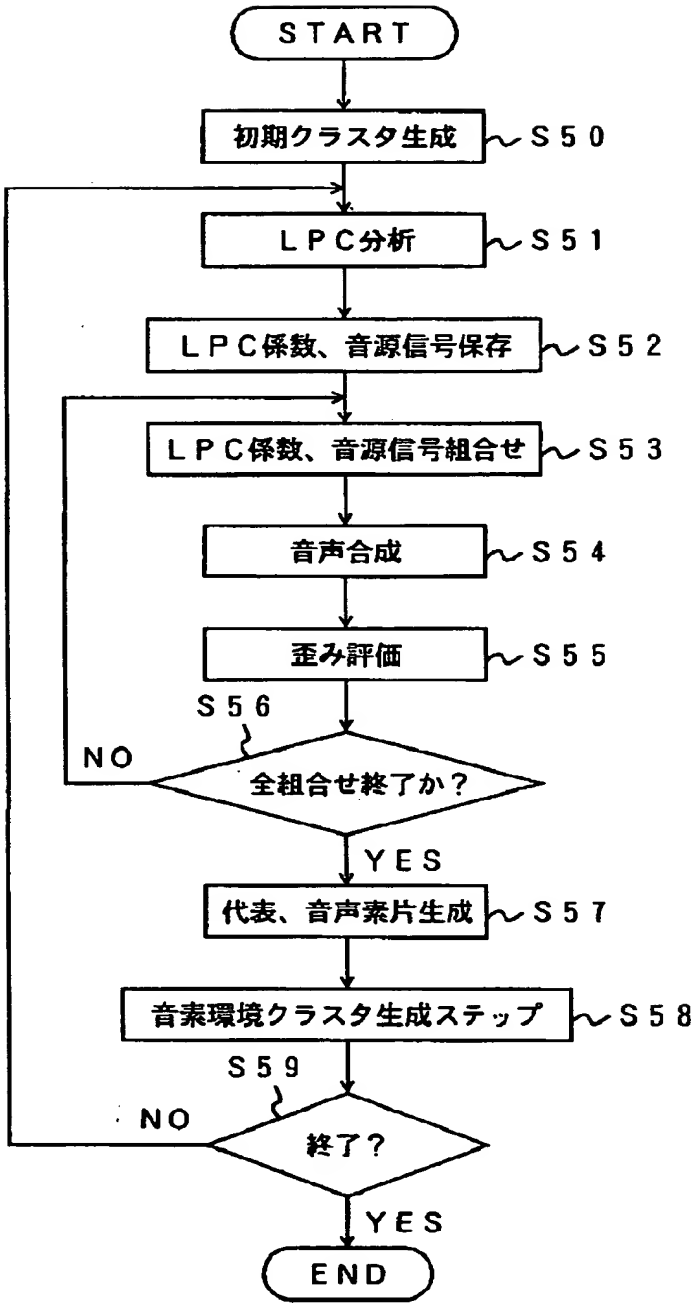
【図１５】


１０２
１０１
音源信号
生成部
合　成
フィルタ
歪
計算部
３４
３６
３１
３２
１０３
ＬＰＣ
分析逆
フィルタ
音源
信号
記憶部
ＬＰＣ
係数
記憶部
３５
３７
３３
最小歪
探索部
クラスタ
記憶部
音素環境
クラスタ
生成部
５２
５１

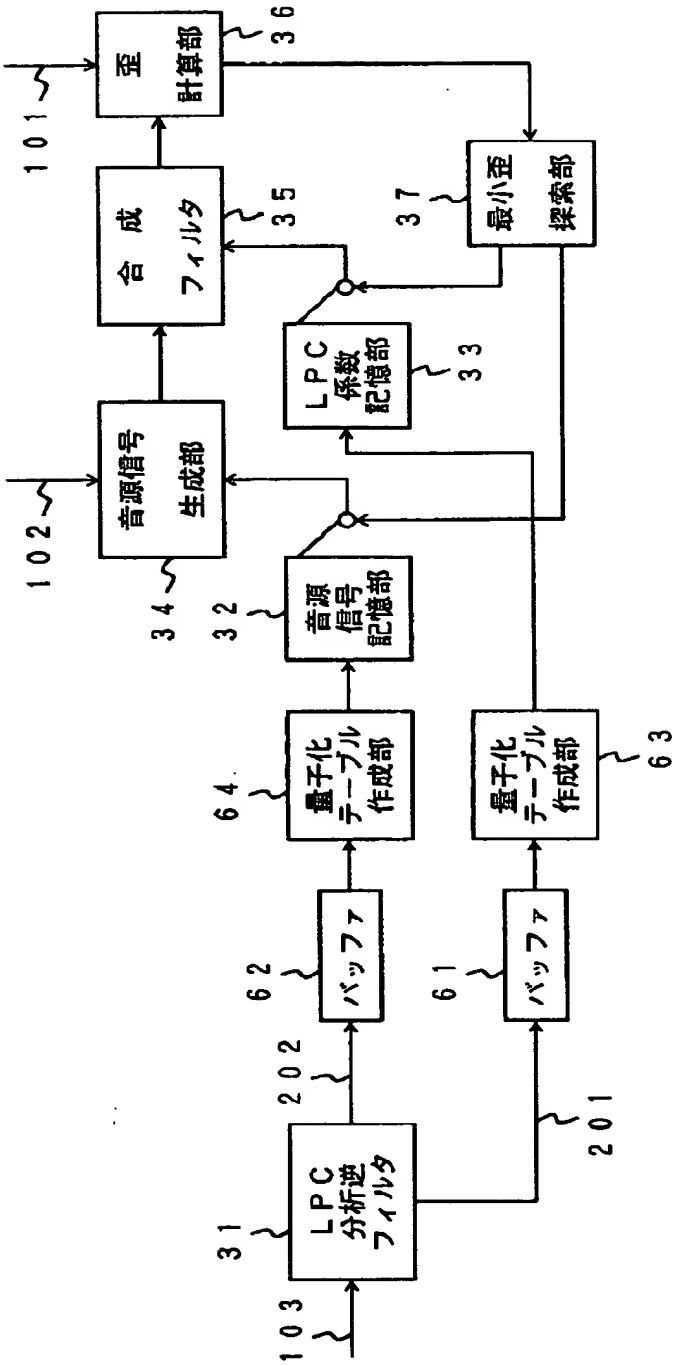
【図１６】


START
ＬＰＣ分析
Ｓ５１
ＬＰＣ係数、音源信号保存
Ｓ５２
ＬＰＣ係数、音源信号組合せ
Ｓ５３
音声合成
Ｓ５４
歪み評価
Ｓ５５
Ｓ５６
NO
全組合せ終了か？
ＹＥＳ
代表、音声素片生成
Ｓ５７
音素環境クラスタ生成ステップ
Ｓ５８
END

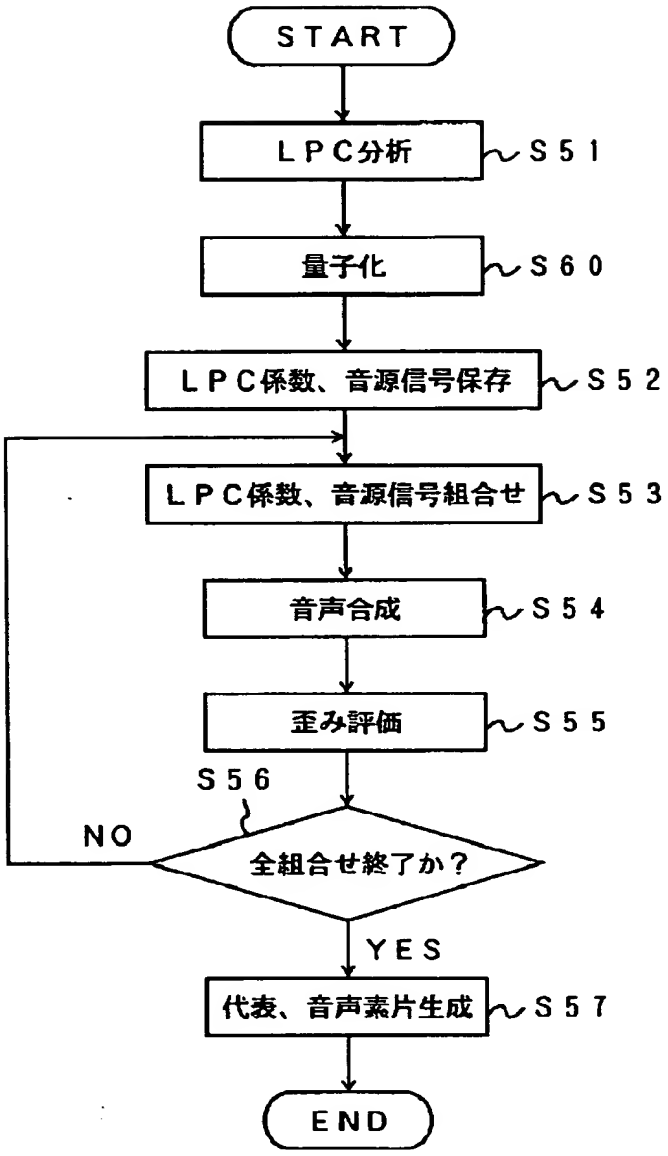
【図１７】


START
初期クラスタ生成
S５０
ＬＰＣ分析
S５１
ＬＰＣ係数、音源信号保存
S５２
ＬＰＣ係数、音源信号組合せ
S５３
音声合成
S５４
歪み評価
S５５
S５６
NO
全組合せ終了か？
ＹＥＳ
代表、音声素片生成
S５７
音素環境クラスタ生成ステップ
S５８
S５９
NO
終了？
ＹＥＳ
END

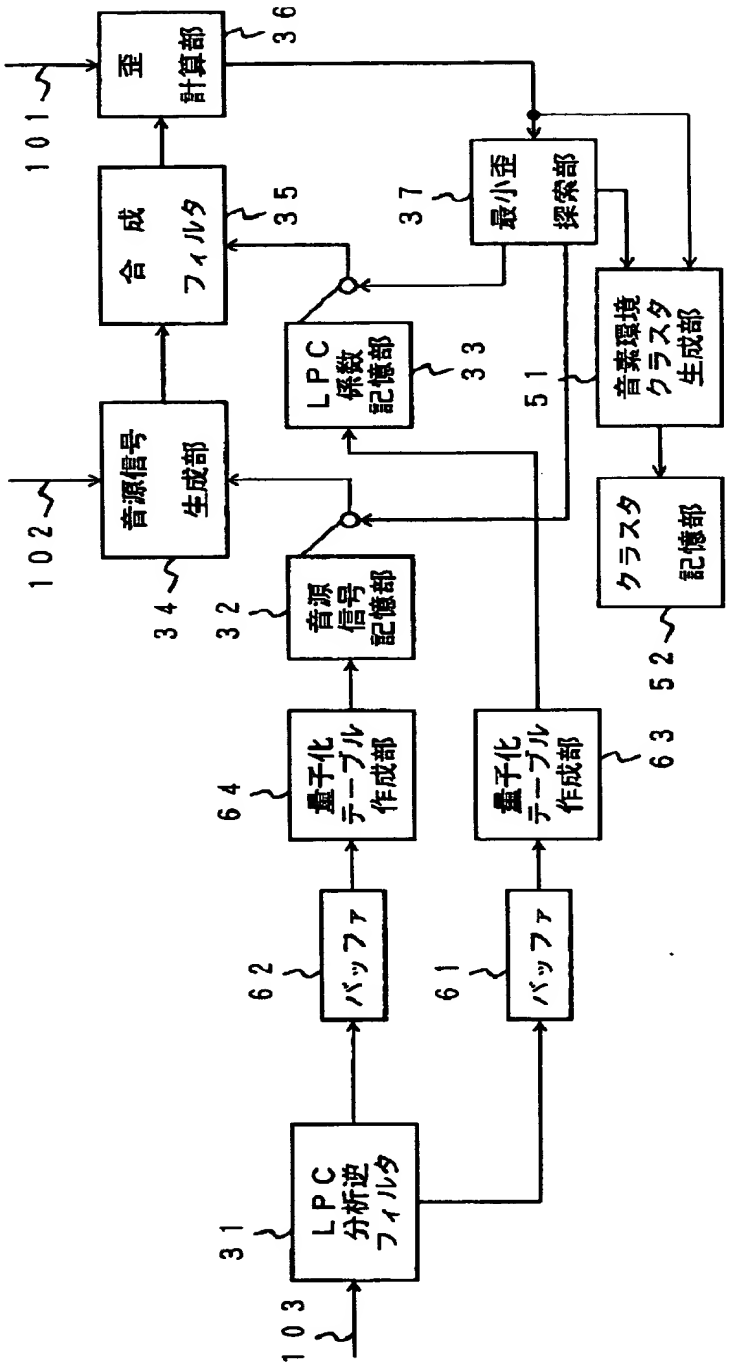
【図１８】


歪 計算部
36
101
合成 フィルタ
35
37
最小歪 探索部
LPC 係数 記憶部
33
音源信号 生成部
102
34
32
音源 信号 記憶部
量子化 テーブル 作成部
64
量子化 テーブル 作成部
63
62
バッファ
61
バッファ
202
201
LPC 分析逆 フィルタ
31
103

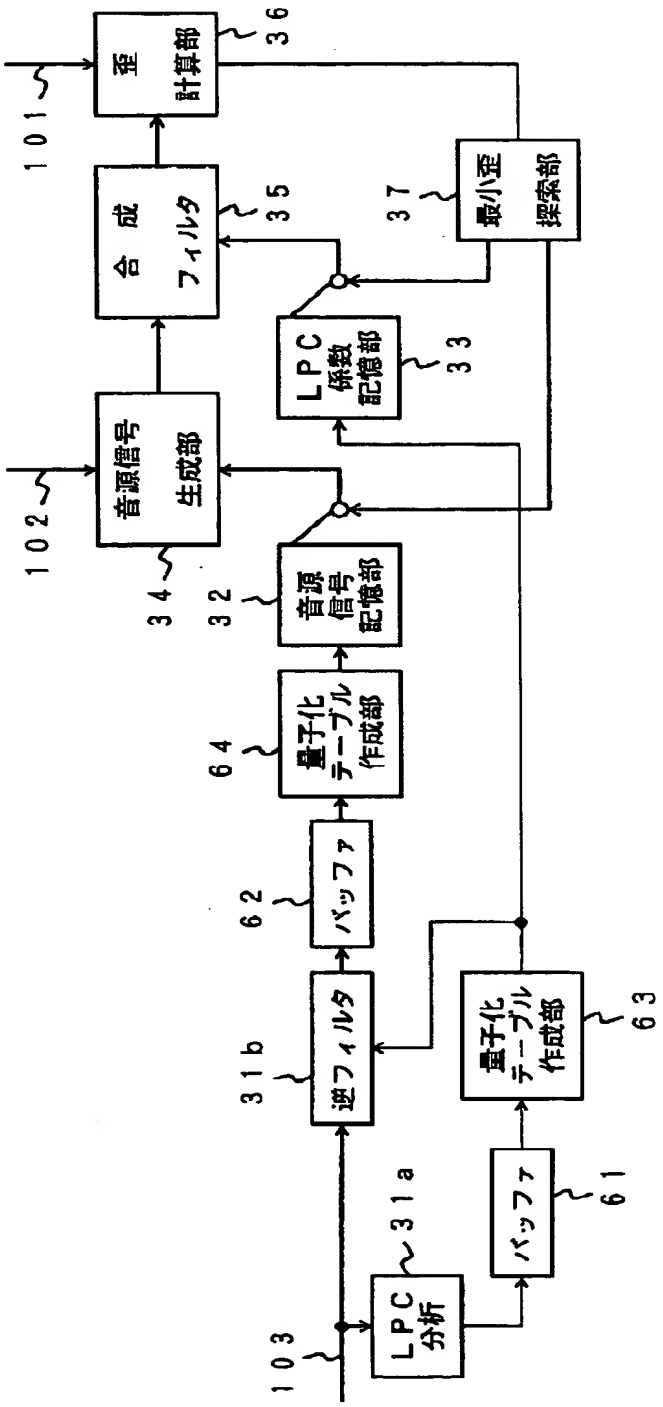
【図１９】


START
ＬＰＣ分析
Ｓ５１
量子化
Ｓ６０
ＬＰＣ係数、音源信号保存
Ｓ５２
ＬＰＣ係数、音源信号組合せ
Ｓ５３
音声合成
Ｓ５４
歪み評価
Ｓ５５
Ｓ５６
NO
全組合せ終了か？
ＹＥＳ
代表、音声素片生成
Ｓ５７
END

【図２０】


歪 計算部
36
101
合成 フィルタ
35
37
最小歪 探索部
LPC 係数 記憶部
33
51
音素環境 クラスタ 生成部
音源信号 生成部
102
34
32
音源 信号 記憶部
クラスタ 記憶部
52
量子化 テーブル 作成部
64
量子化 テーブル 作成部
63
バッファ
62
バッファ
61
LPC 分析逆 フィルタ
31
103

【図２２】


歪 計算部
36
101
合成 フィルタ
35
37
最小歪 探索部
LPC 係数 記憶部
33
音源信号 生成部
102
34
32
音源 信号 記憶部
量子化 テーブル 作成部
64
バッファ
62
逆フィルタ
31b
量子化 テーブル 作成部
63
LPC 分析
31a
バッファ
61
103